length *gag* and *env* nucleotide sequences using the neighbor joining method (see text for details of methodology). Horizontal branch lengths are drawn to scale; vertical separation is for clarity only. Values at the nodes indicate the percent bootstraps in which the cluster to the right was supported (bootstrap values of 75% and higher are shown). Asterisks denote hybrid genomes as determined by additional analyses. Brackets at the right represent the major sequence subtypes of HIV-1 group M. Trees were rooted by using SIVcpzGAB as an outgroup.--

Please replace the paragraph beginning at page 7, line 12, with the following rewritten paragraph:

--Figures 2A-2J. Diversity plots comparing the sequence relationships of the 11 viral genomes described in this patent application to each other and to reference sequences from the database. In each of Figures 2A-2J, the sequence named above the plots is compared to the sequences listed at the right. U455, LAI, C2220, and NDK are published reference sequences for subtypes, A, B, C and D, respectively. Distance values were calculated for a window of 500 bp moved in steps of 10 nucleotides. The x-axis indicates the nucleotide positions along the alignment (gaps were stripped and removed from the alignment). The positions of the start codons of the *gag, pol, vif, vpr, env*, and *nef* genes are shown. The y-axis denotes the distance between the viruses compared (0.05 = 5% divergence).--

Please replace the paragraph beginning at page 7, line 22, with the following rewritten paragraph:

--Figures 3A-3I. Exploratory tree analysis. Neighbor joining trees were constructed for a 500 bp window moved in increments of 100 bp along the multiple genome alignment. Trees depicting discordant branching orders among four of the 11 sequences included in this patent application are shown in Figures 3A-3I (hybrid sequences are boxed). The position of each tree in the alignment is indicated; subtypes are indentified by brackets. Numbers at nodes indicate the percentage of bootstrap values with which the adjacent cluster is supported (only values above 80% are shown). Branch lengths are drawn to scale.--

H3

Please replace the paragraph beginning at page 7, line 30, with the following rewritten paragraph:

--Figures 4A and 4B. Recombination breakpoint analysis for 92W009.6 and 93BR029.4. (Figure 4A) Bootstrap plots depicting the relationship of 92RW009.6 to representatives of subtype A and C, respectively. Trees were constructed from the multiple genome alignment and the magnitude of the bootstrap value supporting the clustering of 92RW009.6 with U455 and 92UG037.1 (subtype A), or C2220 and 92BR025.8 (subtype C), respectively, was plotted for a window of 500 bp moved in increments of 10 bp along the alignment. Regions of subtype A or C origin are identified by very high bootstrap values (>90%). Points of cross-over of the two curves indicate recombination breakpoints. The beginning of *gag, pol, vif, vpr, env,* and *nef* open reading frames are shown. The y-axis indicates the percent bootstrap replicates, which support the clustering of 92RW009.6 with representatives of the respective subtypes. (Figure 4B) Bootstrap plots depicting the relationship of

H4

93BR029.4 to representatives of subtype B and F, respectively. Analyses are as in (Figure 4A), except that bootstrap values supporting the clustering of 93BR029.4 with SF2, OYI, MN, LAI and RF (subtype B), or 93BR020.1 (subtype F), respectively, were plotted. Subtype D viruses were excluded from this analysis because of their known close relationship with subtype B viruses.--

H4 CONT

Please replace the paragraph beginning at page 8, line 16, with the following rewritten paragraph:

--Figures 5A-5D. Recombination breakpoint analysis of 92NG083.2 and 92NG003.1. Neighbor joining trees discordant branching orders of 92NG003.1 and 92NG083.2 in regions delineated by breakpoints identified by distance plots (not shown) are shown in Figures 5A-5D (hybrid sequences are boxed). The position of each tree in the alignment is indicated; subtypes are identified by brackets. Numbers at nodes indicate the percentage of bootstrap values with which the adjacent cluster is supported (only values above 80% are shown). Branch lengths are drawn to scale.--

H5

Please replace the paragraph beginning at page 8, line 27, with the following rewritten paragraph:

--Figures 7A-7C. Subtype specific genome features. (Figure 7A) Alignment of deduced Tat (region encoded by second exon) amino acid sequences. Consensus sequences were generated for available representatitves of all major subtypes (question marks indicate sites at which fewer than 50% of the viruses contain the same amino acid residue). Dashes denote sequence identity with the consensus sequence, while dots represent gaps introduced to

H6

optimize alignments. A vertical box highlights a premature Tat protein truncation (asterisk) which is present in 11 of 15 subtype D, and 4 of 52 subtype B viruses (frequencies are listed in the column on the right). (Figure 7B) Alignment of deduced Rev (region encoded by the second exon) protein sequences. (Figure 7C) Alignment of deduced Vpu protein sequences.--

H6 cont

Please replace the paragraph beginning at page 9, line 19, with the following rewritten paragraph:

--Figures 10A-10K. Exploratory tree analysis. Neighbor joining trees were constructed for a 400 bp window moved in increments of 10 bp along the multiple genome alignment. Trees in Figures 10A-10K depict the discordant branching orders for 94CY032.3 (highlighted). The position of each tree in the alignment is indicated; subtypes are identified by brackets. Numbers at nodes indicate the percentage of bootstrap values with which the adjacent cluster is supported (only values above 80% are shown). Branch lengths are drawn to scale.--

H7

Please replace the paragraph beginning at page 10, line 16, with the following rewritten paragraph:

--Figures 13A-13Z. Nucleotide sequence alignment of the 11 near full-length HIV-1 sequences included in this patent application. Sequences were aligned using CLUSTAL W and adjusted manually using the sequence editor MASE. Dots indicate gaps introduced to optimize the alignment. The beginning and end of all open reading frames are indicated by arrows above or below the alignment. The homologies between the sequences of nucleotides in

H8

H8 Cont'd the eleven independent clones are indicated by dashes. Sequences of nucleotides present uniquely in the various clones (as compared to the corresponding sequences of the other ten clones) are indicated by letters, i.e., the sequences themselves.--

Please replace the paragraph beginning at page 10, line 25, with the following rewritten paragraph:

H9 --Figures 14A and 14B. Amino acid sequence alignments of the Gag polypeptides encoded by the 11 near full-length HIV-1 sequences included in this patent application. The homologies between the sequences of amino acids in the various polypeptides encoded by the eleven independent clones are indicated by dashes. Sequences of amino acids present uniguel in the various polypeptides (as compared to the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., the sequences themselves.--

Please replace the paragraph beginning at page 11, line 1, with the following rewritten paragraph:

H10 --Figures 15A-15C. Amino acid sequence alignments of the Pol polypeptides encoded by the 11 near full-length HIV-1 sequences included in this patent application. The homologies between the sequences of amino acids in the various polypeptides encoded by the eleven independent clones are indicated by dashes. Sequences of amino acids present uniquely in the various polypeptides (as compared to the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., the sequences themselves.--

Please replace the paragraph beginning at page 12, line 12, with the following rewritten paragraph:

--Figures 21A-21C. Amino acid sequence alignments of the Env polypeptides encoded by the 11 near full-length HIV-1 sequences included in this patent application. The homologies between the sequences of amino acids in the various polypeptides encoded by the eleven independent clones are indicated by dashes. Sequences of amino acids present uniquely in the various polypeptide (as compared to the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., the sequences themselves.--

H11

Please replace the paragraph beginning at page 12, line 27, with the following rewritten paragraph:

--The present invention relates to the determination of the nucleic acid sequences of the complete or near complete genomes of 11 non-subtype B HIV-1 viruses isolated from primary isolates collected at major epicenters of the global AIDS pandemic. The nucleotide sequences of these 11 viruses are shown in Figures 13A-13Z (SEQ ID NOS: __ to __).--

H12

Please replace the paragraph beginning at page 14, line 26, with the following rewritten paragraph:

--The present invention relates to nucleic acids having the genomic sequence of any one of the 11 molecular clones for non-subtype HIV-1 isolates of this invention as shown in Figures 13A-13Z (SEQ ID NOS: __ to __), as well as fragments (or partial sequences) thereof. The invention also relates to nucleic acids having

H13

complementary (or antisense) sequences to the sequences shown in Figures 13A-13Z (SEQ ID NOS: __ to __), as well as fragments (or partial sequences) thereof. Partial sequences may be obtained by various methods, including restriction digestion of nucleic acids with sequences shown in Figures 13A-13Z (SEQ ID NOS: __ to __), PCR amplification, and direct synthesis. Partial sequences may be all or part of the LTR and/or other untranslated regions of the genomes of one or more of the 11 viral clones of this invention, and/or all or part of the genes encoding the Gag, Pol, Vif, Vpr, Env, Tat, Rev, Nef and Vpu proteins and/or complementary (or antisense) sequences thereof. Nucleic acids of the invention also include cDNA, mRNA, and other nucleic acids derived from the genomic sequences of one or more of these 11 HIV-1 clones. Sequences of the genes encoding Gag, Pol, Vif, Vpr, Env, Tat, Rev, Nef and Vpu are identified in Figures 13A-13Z.--

Please replace the paragraph beginning at page 16, line 18, with the following rewritten paragraph:

--The nucleic acid probes used in the detection methods set forth above are derived from nucleic acid sequences shown in Figures 13A-13Z (SEQ ID NOS: __ to__). The size of such probes is at least 10-12 bases long, more usually at least about 19 bases long, more usually from about 200 to about 500 bases, and often exceeding about 1000 bases.--

Please replace the paragraph beginning at page 17, line 7, with the following rewritten paragraph:

--The nucleic acid probes used in the detection methods set forth above are derived from sequences substantially homologous to one or more of the sequences shown in Figures 13A-13Z (SEQ ID NOS: __ to __), or their complementary sequences. By "substantially homologous", as used throughout the specification and claims to describe the nucleic acid sequence of the present invention, is meant a high level of homology between the nucleic acid sequence and one or more of the sequences of Figures 13A-13Z (SEQ ID NOS: __ to __), or its complementary sequence. Preferably, the level of homology is in excess of 80%, more preferably in excess of 90%, with a preferred nucleic acid sequence being in excess of 95% homologous with a portion of one or more of the sequences shown in Figures 13A-13Z (SEQ ID NOS: __ to __), or its complement. The size of such probes is usually at least 20 nucleotides, more usually from about 200 to 500 nucleotides, and often exceeding 1000 nucleotides.--

Please replace the paragraph beginning at page 17, line 28, with the following rewritten paragraph:

--The methods for analyzing the RNA for the presence of the viruses of this invention include Northern blotting (94), dot and slot hybridization, filter hybridization (95), Rnase protection (93), and reverse-transcription polymerase chain reaction (RT-PCR)(96). A preferred method is RT-PCR. In this method, the RNA can be reverse transcribed to first strand cDNA using a nucleic acid primer or primers derived from one or more of the nucleotide sequences shown in Figures 13A-13Z (SEQ ID NOS: __ to __). Once the cDNAs are synthesized, PCR amplification is carried out using pairs of primers designed to hybridize with sequences in the genomes of one or more

of the non-subtype B HIV-1 viruses of this invention which are an appropriate distance apart (at least about 50 bases) to permit amplification of the cDNA and subsequent detection of the amplification product. Each primer of a pair is a single-stranded nucleic acid of about 20 to about 60 bases in length where one primer (the "upstream" primer) is complementary to the original RNA and the second primer (the "downstream" primer) is complementary to the first strand of cDNA generated by reverse transcriptions of the RNA. The target sequence is generally about 100 to about 300 bases in length but can be as large as 500-1500 bases or more, e.g., 9,000 bases. Optimization of the amplification reaction to obtain sufficiently specific hybridization to the nucleotide sequences of these viruses is well within the skill in the art and is preferably achieved by adjusting the annealing temperature.--

Please replace the paragraph beginning at page 21, line 27, with the following rewritten paragraph:

--The polypeptides of this invention consist of at least 6-12 amino acids, more preferably at least 3-18 amino acids, even more preferably at least 19-24 amino acids and most preferably at least 25-30 amino acids encoded by, or otherwise derived from, any one of the genomic sequences shown in Figures 13A-13Z (SEQ ID NOS: __ to __).--

Please replace the paragraph beginning at page 31, line 9, with the following rewritten paragraph:

--The present invention further relates to computer-generated alignments of any or more of the nucleotide sequences

shown in Figures 13A-13Z (SEQ ID NOS: __ to __). Computer analysis of the nucleotide sequences, such as the one shown in Figure 13A-13Z, can be carried out using commercially available computer program known to one skilled in the art.--

Please replace the paragraph beginning at page 31, line 14, with the following rewritten paragraph:

--In one embodiment, the sequences shown in Figures 13A-13Z (SEQ ID NOS: __ to __) are aligned by the computer program CLUSTAL (67) and adjusted with multiple-aligned sequence editor (12). The computer analysis results in the distribution of 11 sequences into various genotypes. Five of these sequences represent non-recombinant members of HIV-1 subtypes, and the other six sequences represent HIV-1 intersubtype recombinants.--

Please replace the paragraph beginning at page 32, line 18, with the following rewritten paragraph:

--The multiple computer-generated alignments of nucleotide sequences are shown in Figures 13A-13Z. The multiple computer-generated alignments of encoded amino acid sequences are shown in Figures 14-22. These alignments serve to highlight regions of homology and non-homology between different sequences and hence, can be used by one skilled in the art to design oligonucleotides and polypeptides useful as reagents in diagnostic assays for HIV-1.--

Please replace the paragraph beginning at page 41, line 3, with the following rewritten paragraph:

--To determine the phylogenetic relationships of the viruses described herein, evolutionary trees from full length *gag* and *env* sequences were first constructed. This was done to confirm the authenticity of previously characterized strains, classify the new viruses, and compare viral branching orders in trees from two genomic regions. The results confirmed a broad subtype representation among the selected viruses (Figures 1A and 1B). Strains fell into six of the seven major (non-B) clades, including three for which full length sequences are not available (i.e., F, G and H). However, comparison of the *gag* and *env* topologies also identified three strains with discordant branching orders. 92RW009.6 grouped with subtype C viruses in *gag*, but with subtype A viruses in *env*. Similarly, 93BR029.4 clustered with subtype B viruses in *gag*, but with subtype F viruses in *env*. 94CY017.41 appeared to cluster within subtype A viruses in *env*, but fell into an unknown subtype in *gag*. However, characterization of the latter strains is still ongoing. These different phylogenetic positions were supported by high bootstrap values and thus indicated that these strains were intersubtype recombinants.--

Please replace the pargraph beginning at page 43, line 27, with the following rewritten paragraph:

--To examine the phylogenetic position of the newly derived strains relative to each other and to the reference sequences over the entire genome, exploratory tree analyses were performed using the same multiple genome alignment generated for the diversity plots (Figures 3A-3I). A total of 79 trees were constructed for overlapping fragments of 500 bp, moved in 100 bp increments

along the alignment. As expected, four genomes were identified that clustered in different subtypes in different parts of their genome. These included 93BR029.4 which alternated between subtypes F and B, 92RW009.6 which alternated between subtypes A and C, and 92NG083.2 and 92NG003.1 which grouped either independently or within subtype A. Interestingly, the latter two strains exhibited distinct patterns of mosaicism. In trees spanning the region 3501-4000, 92NG003.1 clustered within subtype A, while 92NG083.2 clustered independently, presumably representing subtype G. In contrast to these strains, there was no evidence for a hybrid genome structure in 94IN476.104, 96ZM651.8, 96ZM751.3, 93BR020.1 or 90CF056.1. These viruses branched consistently in all regions analyzed. Based on these findings and the results from the diversity plots, it appeared that five of the eleven selected HIV-1 strains represent non-recombinant reference strain for subtypes C (94IN476.104, 96ZM651.8, 96ZM751.3), F (96BR020.1) and H (90CF056.1), respectively, while at least five are intersubtype recombinants. CY017.41 may be recombinant, but work is in progress in this regard.--

Please replace the paragraph beginning at page 44, line 19, with the following rewritten paragraph:

--To map the location of the recombination breakpoints in 92RW009.6 and 93BR029.4, bootstrap plots and informative site analyses were used (18,52,23). Unrooted trees were constructed which included U455, 92UG037.1, LAI, MN, OYI, SF2, RF, C2220, 92BR025.1, NDK, ELI, Z2Z6, 93BR020.1 and 90CF056.1; then the magnitude of the bootstrap values supporting (i) the clustering of

92RW009.6 with members of subtype A (U455, 92UG037.1) or C (2220, 92BR025.8), as well as (ii) the clustering of 93BR029.4 with members of subtype B (LAI, MN, OYI, MN, RF) or F (92BR020.1) was determined (in the latter case subtype D viruses were excluded because of their known close relationship to subtype B viruses). Figures 4A and 4B depict the results of 797 such phylogenetic analyses generated for each genome, performed on a window of 500 nucleotides moved in steps of 10 nucleotides. Very high bootstrap values (>80%) supporting the clustering of 92RW009.6 with subtype C were apparent in *gag*, the 3' two-thirds of *pol*, and *nef*. By contrast, significant branching of 92RW009.6 with subtype A was apparent in the *gag/pol* overlap and the *env* region. In a small region (4,000 to 4,200) in the middle of the genome, 92RW009.6 appeared not to cluster significantly with either subtype, but further inspection revealed that this was due to a small number of informative sites. These data thus indicated four points of recombination crossovers between subtypes A and C (Figure 4A). A similar analysis identified six recombinant breakpoints between subtypes B and F in 93BR029.4 (Figure 4B). Theses included two more (in *gag*) than were apparent from the diversity plot analysis (compare Figures 2A-2J), indicating a greater sensitivity of this approach.--

Please replace the paragraph beginning at page 46, line 21, with the following rewritten paragraph:

--Because of the lack of a full length subtype G reference sequence, recombination breakpoint analysis of 92NG003.1 and 92NG083.2 required a different approach. The analyses summarized in Figures 2A-2J and Figures 3A-3I suggested that these two viruses

contained subtype A sequences in the middle of their genome. To attempt to confirm this, and to define the extent of these putative subtype A fragments, a more detailed diversity plot analysis of the viral middle region (between position 3,000 and 6,000) was performed using different viral strains and varying window sizes (ranging from 200 to 400 bp) to examine the extent of sequence divergence of 92NG083.2 and 92NG003.1 from members of other subtypes, including subtype A. Diversity plots for 92NG003.1 compared to U455, C2220, NDK and 92NG083.2 and for 92NG083.2 compared to U455, C2220, NDK and 92NG003.1 depicted representative results (using a window size of 300 bp moved in steps of 10 bp along the alignment)(data not shown). Similar to the data shown in Figures 2A-2J, the two "subtype G" viruses are roughtly equidistantly related to members of subtypes A (U455), C(C2220), and D(NDK), except for two regions in 92NG003.1 and one region in 92NG083.2 where both viruses are disproportionately more closely related to U455 than they are to each other. Noting the points at which the "G"-A distance increases or decreases relative to the others allowed the tentative identification of recombination breakpoints. For example, at position 3400, the U455 plot falls whereas the C220, NDK and 92NG083.2 plots do not, and around site 3600 the U455 plot crosses the 92NG083.2 plot. Bearing in mind the window size of 300 nucleotides, this finding suggested that a recombination cross-over occurred around position 3500. Similar "G"-A plot crossing around positions 3800, 4200 and 5200 (in the diversity plot for 92NG003.1), and around positions 4200 and 4800 (in the diversity plot for 92NG083.2), suggested additional recombination breakpoints.--

Please replace the paragraph beginning at page 47, line 15, with the following rewritten paragraph:

--Phylogenetic trees were then constructed using the regions of sequence defined by these putative breakpoints (Figures 5A-5D). This analysis generally supported the conclusions drawn from the diversity plots, i.e., 92NG003.1 clustered with subtype A viruses in the region between 3501 and 3800, whereas 92NG083.2 did not; and both 92NG003.1 and 92NG083.2 clustered with subtype A viruses in the region 4201 and 4800. However, neither the diversity plot nor the tree analysis allowed the definition of the boundaries of the subtype A fragments with certainty. Nevertheless, the data indicated that (i) both 92NG083.2 and 92NG003.1 represent G/A recombinants, (ii) that they are the result of different recombination events because some of their breakpoints are clearly different, and (iii) that 92NG083.2 likely encodes a non-recombinant *pol* gene. A schematic representation of the mosaic genomes of 92NG083.2 and 92NG003.1 is shown in Figure 6.

H26

Please replace the paragraph beginning at page 47, line 29, with the following rewritten paragraph:

--Having classified the new viruses with respect to their subtype assignments, their sequences were examined for clade-specfic signature sequences. Comparing deduced amino acid sequences gene by gene, several subtype specific features were found (Figures 7A-7C). For example, most subtype D viruses contain an in-frame stop codon in the second exon of *tat*, which removes 13 to 16 amino acids from the carboxy terminus of the Tat protein

H27

(Figure 7A). Similarly, all subtype C viruses (including 94IN476.104, 96ZM651.8 and 96ZM751.3) contain a stop codon in the second exon of *rev* which would be predicted to shorten their protein by 16 amino acids (Figure 7B). Subtype C viruses also contain a 15 base pair insertion at the 5' end of the *vpu* gene (Figure 7C) which extends the putative membrane spanning domain of the Vpu protein by 5 amino acids (data not shown). Although these changes are unlikely to alter the function of the respective gene products in a major way (e.g., the known functional domains of both Tat and Rev proteins are not affected by these changes), it is possible that they could influence their mechanism of action in a subtle (but nevertheless biologically important) manner.--

H27 Cont

Please replace the paragraph beginning at page 54, line 4, with the following rewritten paragraph:

--To map potential recombination breakpoints in this remaining region, four recently reported, partial but non-mosaic subtype G sequences from Mali which spanned the *vif/vpr* region and thus bridged the "subtype A gap" or 92NG083.2 were used (77). A set of distance plots that compare 94CY032.3 to one of these newly derived G sequences (95ML045) as well as representatives of subtype A (U455), B(MN), and D(ELI), respectively, were constructed (data not shown). Consistent with the results from the exploratory tree analysis (Figures 4A and 4B), 94CY032.3 was disproportionately more closely related to U455 in the 5' and 3' thirds of this fragment, suggesting the presence of subtype A-like segments. However, in the middle of the fragment, 94CY032.3 was clearly equidistant from U455 and the other subtypes, suggesting an independent position

H28

(diversity plots were generated for a window of 300 bp moved in increments of 10 bp). Thus, noting the points at which the "A" distance increased and decreased relative to the other distances allowed us to tentatively map the two remaining breakpoints, one at 4650 and the other at 5000. Trees constructed from sequences surrounding these two breakpoints (Figure 12) confirmed that 94CY032.3 switched position from subtype A (Figure 12; panel 4255-650) to subtype I (panel 4651-5000), and back to subtype A (5001-5300; note, that the new subtype G sequences only cover the region between 4255 and 5300).--

#28 Cont.